EmoGest: Investigating the Impact of Emotions on Spontaneous Co-speech Gestures

Kirsten Bergmann[1], Ronald Böck[2], Petra Jaecks[3]
[1]Faculty of Technology, Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
[2]Cognitive Systems Group, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany
[3]Faculty of Linguistics and Literary Studies, Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
kirsten.bergmann@uni-bielefeld.de, ronald.boeck@ovgu.de, petra.jaecks@uni-bielefeld.de

Abstract

Spontaneous co-speech gestures are an integral part of human communicative behavior. Little is known, however, about how they reflect a speaker's emotional state. In this paper, we describe the setup of a novel body movement database. 32 participants were primed with emotions (happy, sad, neutral) by listening to selected music pieces and, subsequently, fulfilled a gesture-eliciting task. We present our methodology of evaluating the effects of emotion priming with standardized questionnaires, and via automatic emotion recognition of the speech signal. First results suggest that emotional priming was successful, thus, paving the way for further analyses comparing the gestural behavior across the three experimental conditions.

Keywords: Emotions, priming, multimodal, co-speech gestures, corpus collection, audio feature analysis

1. Introduction

There is a large body of empirical evidence demonstrating that emotional states manifest themselves in different aspects of communicative behavior. For speech, research has demonstrated various effects in terms of acoustic features such as loudness, speaking rate, intonation, voice quality etc., as well as lexical choice, use of syntactic construction etc. (see, e.g., Bänziger et al. (2014)). Likewise, facial expressions have been studied extensively as a major medium of expressing emotions (see, e.g., Keltner et al. (2003), Russell et al. (2003)). In addition, there is a substantial amount of evidence demonstrating that particular body postures are associated with a specific mood or attitude (e.g., Crane and Gross (2013), Dael et al. (2012)).

Despite all this, what we know about the impact of particular emotional states on *co-speech gestures* is still sparse. Existing corpora like the *Belfast database* (Douglas-Cowie et al., 2000), the *EmoTV* corpus (Abrilian et al., 2005), or the *GEMEP* corpus (Bänziger et al., 2010) do not focus on co-speech gestures in detail. There are, however, a few studies which have begun to address the impact of emotional states on speech-accompanying gestures. Castellano et al. (2007) conducted a study in which participants performed one and the same gesture while expressing different emotional conditions. An approach of automated video analysis has been employed to investigate whether expressive motion cues, such as movement amplitude or speed/fluidity of movement, allow to discriminate between emotions. Results showed that expressive motion cues allow to discriminate between high and low arousal emotions as well as positive and negative emotions. Kipp and Martin (2009) investigated how basic gesture form features (handedness, hand shape, palm orientation, motion direction) are related to components of emotion. The analysis was based on a corpus of segments from two versions of a movie in which the protagonist displays a wide range of emotions. The analysis revealed that handedness in gestures is closely correlated with emotion categories. A positive correlation was demonstrated for high pleasure and left-handed gestures, while right-handed gestures were more likely to occur when low pleasure was expressed. With a similar approach, Fourati and Pelachaud (2013) recently set up a larger database of acted emotional body behavior. 3D motion capture data synchronized with full HD video was recorded from 11 actors who expressed different emotional states while describing several actions. In advance, the actors had gone through a training to express emotions in daily actions while avoiding exaggerated and expressiveless behavior.

The present corpus collection aims to advance this previous work by providing detailed data on the interrelation of emotions and co-speech gestures in spontaneous face-to-face interaction. While the aforementioned studies took important steps in providing first data and evidence that different aspects of gesture use are affected by the speakers' emotional state, they are limited to *acted* emotional states. The question, therefore, remains whether and how *spontaneous* speech-accompanying gestures reflect the speaker's emotional state. Likewise, in the community of speech-based emotion recognition, there is a recent trend towards naturalistic data sets which represent spontaneous emotional reactions (see, e.g., Schuller et al. (2011)).

In this paper, we describe the setup of a novel database of spontaneous co-speech body movement behavior, the *EmoGest corpus*. Participants were primed with emotions by listening to selected music pieces – rather than instructed to express particular emotions – and subsequently fulfilled a gesture-eliciting task. In the following we will first sketch the study setup. Then, we put a focus on our methodology and first results of evaluating the effects of emotional priming in terms of (a) participants' self-ratings with standardized questionnaires as well as (b) automatic emotion recognition of the speech signal. We conclude with a prospect of gesture coding techniques intended to complement the corpus data.

2. Experimental Setup and Data Collection

The corpus was set up based on a linguistic experiment. 32 participants interacted naturally in a tangram task, where they had to describe 12 tangram figures to a confederate interaction partner. Prior to the tangram task, all participants listened to one of three audio files of about three minutes length each presenting classical musical pieces that induce different emotions (happiness, sadness, neutral). The happy and sad stimuli were collected and published by Eerola and Vuoskoski (2011). The items of their "Soundtracks datasets for music and emotion" were evaluated for their power to induce emotions (see Eerola and Vuoskoski (2011) for statistics). The neutral stimuli were generated according to the description and statistics by Hunter et al. (2008). After participants were provided with the music stimuli, they completed self-rating questionnaires to evaluate the priming effect of the musical emotion induction. Subsequently, they listened to the same music stimulus once again before they fulfilled the tangram description task in interaction with a confederate.

The primary data of the corpus consists of audio and HD video recordings of the interactions as well as Kinect data. For the videotape three synchronized camera views were recorded (see Fig. 1). In total, the corpus consists of ∼12 hours of dialogical interaction and contains ∼4.000 representative gestures (projected from first gesture segmentations of ∼25% of the material). The three experimental groups were comparable in handedness according to the Edinburgh handedness inventory ((Oldfield, 1971); 27 right, 4 left, 1 ambidextrous; χ^2=2.651, p=0.618) and gender distribution (χ^2=3.269, p=0.195). They did not differ in age (20-41 years, χ^2=2.327, p=0.312) or years of education (13-25 years, χ^2=1.420, p=0.492).

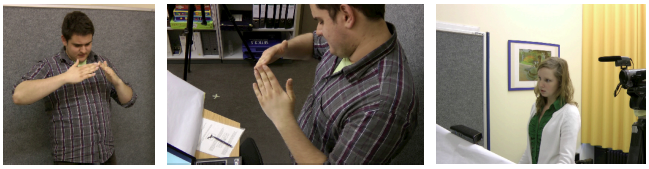

Figure 1: Experimental dialogue situation from three camera views, capturing a participant who describes a stimulus tangram figure displayed on a laptop (left and middle), and the confederate (right).

Several personality questionnaires were conducted (prior to the main experiment). There were no significant differences in personality traits across the three groups (BFI-K, Rammstedt and John (2005); e.g. extraversion: χ^2=4.409, p=0.110), actual mood (UWIST, Matthews et al. (1990); χ^2=0.384, p=0.825) or empathy (SPF/IRI, Paulus (2009); χ^2=0.670, p=0.715).

3. Evaluation of Emotional Priming

3.1. Self-ratings of Emotional State

To evaluate the priming effect of the musical emotion induction, two different scales were applied. After listening to the music, the groups differed in their feelings of 'joyful activation', 'wonder', 'power', 'tension', 'sadness' (GEM Scales, (Zentner et al., 2008)) and valence and activity (dimensional model, Eerola and Vuoskoski (2011)). For example, 'joyful activation' is rated significantly higher in the 'happy' condition (χ^2=16.474, p<.001) providing evidence for a relevant emotional priming effect. Therefore, we argue that it is scientifically sound to compare the three condition groups in further analyses.

3.2. Analysis of Audio Features

To complement the results from participants' self estimation of their emotional state, we employ an automated analysis of acoustic features. In the field of speech data-based emotion recognition two categories of features are widely used, namely spectral and prosodic features. Most speech recognition systems rely on spectral features sets which are based on Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coefficients (LPCs), and Perceptual Linear Predictive coefficients (PLPs). Various researchers showed that these features are also suitable to recognize emotions from speech (cf. Böck et al. (2010), Schuller et al. (2010), Ververidis and Kotropoulos (2006), Vogt and André (2005)). On the other hand, prosodic features like pitch, intensity, voice quality and vocal tract features provide additional information for the classification of emotional speech. Vocal tract features like formants, their bandwidths etc. reflect characteristics of the speaker whereas voice quality features (jitter, shimmer, etc.) characterize the current articulation. Reviews on prosodic features are given in Schuller et al. (2011), Ververidis and Kotropoulos (2006). The most important issue in feature selection is the identification of meaningful features that represent the characteristics of the speaker and the current situation. Especially, in the context of naturalistic interactions existing and well-known feature sets have to be re-evaluated.

The EmoGest corpus provides a naturalistic Human-Human Interaction (HHI) of two partners, the participant and a confederate. The participant was primed to be in a certain emotional state, namely happy or sad (or neutral as a control). To evaluate the priming from a speech perspective we concentrate on the two emotions which can be also classified as positive and negative. From these considerations and based on previous work (Böck et al., 2012), we selected features which will be in the focus of future research: the first to third formant and their corresponding bandwidth, pitch, jitter, and intensity (Scherer, 2001; Vlasenko et al., 2011) are potentially meaningful since these are related to negative as well as high aroused emotions (cf. De Looze et al. (2011), Schuller et al. (2010)).

The feature extraction is conducted on a level of utterances. To extract the features we applied PRAAT (cf. Boersma (2001)) for prosodic features and the Hidden Markov Toolkit (HTK) (cf. Young et al. (2009)) for MFCCs and combined them afterwards. In preliminary tests such a procedure is advisable since the combination of features can be handled more easily.

3.2.1. Classifiers

In the community of emotion recognition from speech several types of classifiers are used whereas Support Vector Machines and Hidden Markov Models (HMMs) are most prominent. HMMs are utilized in the classification of emotional speech (cf. e.g. El Ayadi et al. (2011), Schuller et al. (2011)). In general, each HMM is a finite state automata which passes from state s_i to state s_j in each time slot. While traversing the model a sequence of observations is produced given a certain probability density. Given a set of trained HMMs the most likely sequence of observations is calculated by the Viterbi algorithm. Afterwards, the model providing the highest log-likelihood is selected as the classification result. Further technical details are given in El Ayadi et al. (2011), Young et al. (2009).

Since we are dealing with a multi-modal corpus we have the opportunity to investigate single modalities in the context of naturalistic HHI and further, to combine various modalities. This leads to the issue of fusion. According to (Krell et al., 2013) we suggest a two step classification process. For each modality features are extracted separately and afterwards, are used to achieve a first classification results. This will be finally combined with those results gained by applying the other modalities. To handle gaps in the input sequence of the final classifier., that means, information is partially not available, a suitable combination method has to be identified. As discussed by Krell et al. (2013), Markov Fusion Networks can be a potential solution.

3.2.2. Preliminary results

An automatic emotion recognition from speech was conducted applying HMMs and the feature set described above in a 10-fold-cross-validation. Based on a subset of the data we achieved an unweighted average accuracy of 90.8% in a two class investigation given by the experimental design ('happy' vs. 'sad'). In line with our results from participants' self-rating of their emotional state, these results indicate that the emotional priming was successful and that the speakers' emotional state can be automatically distinguished in speech. As up to now, not all participants of the experiment were processed to enable automatic classification, the presented results do not have high significance, yet. The preliminary study was implemented to verify if the priming could be seen also in emotionally colored speech.

4. Conclusion

Our goal is to provide a corpus which allows to address whether and how *spontaneous* co-speech gesture use in terms of gesture rate, gesture types, physical gesture form, and gesture expressivity (cf. Hartmann et al. (2006)) is affected by emotional states of the speaker. In this paper, we described the experimental setup of the corpus collection and focused on evaluations of the applied emotional priming. First results are promising so that we now continue to set up the full corpus. The audio signal-based evaluation will be continued and further complemented with an observer-based rating of speakers' emotional state. In addition, we will continue to generate secondary data, particularly focusing on speakers' gestural behavior. To this end, we will apply a feature-based coding of physical gesture form as already applied in the SaGA corpus (Lücking et al., 2013) complemented with annotations according to the NEUROGES coding system (Lausberg, 2013). We will further apply automated coding techniques based on Kinect data, e.g., the MINT.tools (Kousidis et al., 2013), or the NovA for social signal analyses (Baur et al., 2013). These codings will enable us to conduct detailed analyses of how spontaneous co-speech gesture use is affected by emotional states, as well as detailed inter-modal analyses of linguistic content, speech, and gestures in emotionally primed speakers.

5. Acknowledgement

We acknowledge support by the Collaborative Research Centre SFB 673 "Alignment in Communication" and the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems", both funded by the German Research Foundation (DFG).

6. References

Abrilian, S., Devillers, S., Buisine, S., and Martin, J.-C. (2005). EmoTV1: Annotations of real-life emotions for the specification of multimodal affective interfaces. In *Human Computer Interaction International*.

Bänziger, T., Scherer, and R., K. (2010). Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In Scherer, K. R., Bänziger, T., and Roesch, E. B., editors, *Blueprint for affective computing: A sourcebook*, pages 271–294. Oxford University Press, Oxford, UK.

Bänziger, T., Patel, S., and Scherer, K. (2014). The role of perceived voice and speech characteristics in vocal emotion communication. *Journal of Nonverbal Behavior*, 38(31–52).

Baur, T., Damian, I., Lingenfelser, F., Wagner, J., and André, E. (2013). Nova: Automated analysis of nonverbal signals in social interactions. In Salah, A. A., Hung, H., Aran, O., and Gunes, H., editors, *Human Behavior Understanding*. Springer, Berlin/Heidelberg.

Böck, R., Hübner, D., and Wendemuth, A. (2010). Determining optimal signal features and parameters for hmm-based emotion classification. In *Proceedings of the 15th IEEE Mediterranean Electrotechnical Conference*, pages 1586–1590, Valletta, Malta. IEEE.

Böck, R., Limbrecht, K., Siegert, I., Glüge, S., Walter, S., and Wendemuth, A. (2012). Combining mimic and prosodic analyses for user disposition classification. In Wolff, M., editor, *Proceedings of the 23. Konferenz Elektronische Sprachsignalverarbeitung*, pages 220–228, Cottbus, Germany.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Castellano, G., Villalba, S., and Camurri, A. (2007). Recognizing human emotions from body movement and gesture dynamics. In Paiva, A., Prada, R., and Picard, R., editors, *Affective Computing and Intelligent Interaction*, LNAI 4738, pages 71–82. Springer, Berlin/Heidelberg.

Crane, E. and Gross, M. (2013). Effort-shape characteristics of emotion-related body movement. *Journal of Nonverbal Behavior*, 37(2):91–105.

Dael, N., Mortillaro, M., and Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, 12(5):1085–1101.

De Looze, C., Oertel, C., Rauzy, S., and Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *17th International Congress of Phonetic Sciences*, Hong Kong, China.

Douglas-Cowie, E., Cowie, R., and Schröder, M. (2000). A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA Workshop on Speech and Emotion*.

Eerola, T. and Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39:18–49.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587.

Fourati, N. and Pelachaud, C. (2013). A new emotional body behavior database. In Edlund, J., Heylen, D., and Paggio, P., editors, *Proceedings of the Workshop on Multimodal Corpora 2013: Multimodal Corpora: Beyond Audio and Video*.

Hartmann, B., Mancini, M., and Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In Gibet, S., Courty, N., and Kamp, J.-F., editors, *Gesture in Human-Computer Interaction and Simulation*, pages 45–55. Springer, Berlin/Heidelberg.

Hunter, P. G., Schellenberg, E. G., and Schimmack, U. (2008). Mixed affective responses to music with conflicting cues. *Cognition & Emotion*, 22(2):327–352.

Keltner, D., Ekman, P., Gonzaga, G., and Beer, J. (2003). Facial expression of emotion. In *Handbook of affective sciences*, pages 415–532. Oxford University Press, New York, NY, US.

Kipp, M. and Martin, J.-C. (2009). Gesture and emotion: Can basic gestural form features discrminate emotions? In Cohn, J., Nijholt, A., and Pantic, M., editors, *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII-09)*. IEEE Press.

Kousidis, S., Pfeiffer, T., and Schlangen, D. (2013). MINT.tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. In *Proceedings of Interspeech 2013*.

Krell, G., Glodek, M., Panning, A., Siegert, I., Michaelis, B., Wendemuth, A., and Schwenker, F. (2013). Fusion of fragmentary classifier decisions for affective state recognition. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, volume 7742 of *LNAI*, pages 116–130. Springer Berlin Heidelberg.

Lausberg, H., editor. (2013). *Understanding body movement: A guide to empirical research on nonverbal behavior: with an introduction to the NEUROGES coding system*. PL Academic Research, Frankfurt a.M.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2013). The bielefeld speech and gesture alignment corpus (SaGA). *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Matthews, G., Jones, D., and Chamberlain, A. (1990). Refining the measurement of mood: The uwist mood adjective checklist. *British Journal of Psychology*, 81:17—42.

Oldfield, R. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9:97–113.

Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen (IRI) zur Messung von Empathie. Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. [The Saarbrücken personality questionnaire (IRI) for measuring empathy: A psychometric evaluation of the German version of the interpersonal reactivity index].

Rammstedt, B. and John, O. P. (2005). Kurzversion des big five inventory (bfi-k): Entwicklung und validierung eines ökonomischen inventars zur erfassung der fünf faktoren der persönlichkeit. *Diagnostika*, 51:195–206.

Russell, J. A., Bachorowski, J., and Fernández-Dols, J. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54:329–349.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research,*, pages 92–120.

Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., and Wendemuth, A. (2011). Vowels formants analysis allows straightforward detection of high arousal emotions. In *2011 IEEE International Conference on Multimedia and Expo (ICME)*, Barcelona, Spain.

Vogt, T. and André, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia and Expo 2005*, pages 474–477, Amsterdam, The Netherlands. IEEE.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2009). *The HTK Book, version 3.4*. Cambridge University Engineering Department.

Zentner, M., Grandjean, D., and Scherer, K. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4):494—521.